LING 575: Seminar on statistical machine translation

Spring 2011
Lecture 3

Kristina Toutanova
MSR & UW

With slides borrowed from Philipp Koehn
Overview

- A bit more on EM for IBM model 1
  - Example on p.92 of book
- Other word-based translation and alignment models
- Phrase-based translation model
  - Phrase extraction
  - Model Extensions
    - Other features and discriminative estimation
    - Order Model
  - Probabilistic models for phrase-based translation
EM for IBM Model 1
EM for IBM 1 example

Ignoring the NULL word in the source for simplicity.
Collecting counts for M-step

- The expected count for word $f$ translating to word $e$ given sentence pair $e, f$:

\[
c(e|f; e, f) = \sum_a p(a|e, f) \sum_{j=1}^{l_e} \delta(e, e_j) \delta(f, f_{a(j)})
\]

- Can be efficiently computed as follows, using similar rearranging:

\[
c(e|f; e, f) = \frac{t(e|f)}{\sum_{i=0}^{l_f} t(e|f_i)} \sum_{j=1}^{l_e} \delta(e, e_j) \sum_{i=0}^{l_f} \delta(f, f_i)
\]
Collecting counts

- \( c(\text{the} | \text{das}; \text{the house, das Haus}) = t(\text{the} | \text{das}) / \{t(\text{the} | \text{das}) + t(\text{the} | \text{Haus})\} = .25 / (.25 + .25) = .5 \)
- \( c(\text{the} | \text{Haus}; \text{the house, das Haus}) = t(\text{the} | \text{Haus}) / \{t(\text{the} | \text{Haus}) + t(\text{the} | \text{das})\} = .5 \)
- \( c(\text{house} | \text{das}; \text{the house, das Haus}) = .5 \)
- \( c(\text{house} | \text{Haus}, \text{the house, das Haus}) = .5 \)
- \( c(\text{the} | \text{das}; \text{the book, das Buch}) = .5 \)
- \( c(\text{the} | \text{Buch}; \text{the book, das Buch}) = .5 \)
- \( c(\text{book} | \text{das}; \text{the book, das Buch}) = .5 \)
- \( c(\text{book} | \text{Buch}; \text{the book, das Buch}) = .5 \)
- ….
Adding up the counts across sentences

- $c(\text{the}|\text{das}) = c(\text{the}|\text{das}; \text{the house, das Haus}) + c(\text{the}|\text{das}; \text{the book, das Buch}) = 1$
- $c(\text{house}|\text{das}) = 0.5$
- $c(\text{book}|\text{das}) = 0.5$
- $c(\text{house}|\text{Haus}) = 0.5$
- $c(\text{the}|\text{Haus}) = 0.5$
- ....
M-step for IBM Model 1

- After collecting counts from all sentence pairs, we add them up and re-normalize to get new lexical translation probabilities:

\[
t(e|f; e, f) = \frac{\sum_{(e,f)} c(e|f; e, f))}{\sum_{f} \sum_{(e,f)} c(e|f; e, f))}
\]

- \(t(\text{the}|\text{das})=1/(1+.5+.5)=.5\)
- \(t(\text{house}|\text{das})=.5/2=.25\)
- \(t(\text{book}|\text{das})=.5/2=.25\)
Parameters at convergence

<table>
<thead>
<tr>
<th>e</th>
<th>f</th>
<th>initial</th>
<th>1st it.</th>
<th>2nd it.</th>
<th>3rd it.</th>
<th>...</th>
<th>final</th>
</tr>
</thead>
<tbody>
<tr>
<td>the</td>
<td>das</td>
<td>0.25</td>
<td>0.5</td>
<td>0.6667</td>
<td>0.7996</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>book</td>
<td>das</td>
<td>0.25</td>
<td>0.25</td>
<td>0.1795</td>
<td>0.1115</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>house</td>
<td>das</td>
<td>0.25</td>
<td>0.25</td>
<td>0.1538</td>
<td>0.0889</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>the</td>
<td>buch</td>
<td>0.25</td>
<td>0.25</td>
<td>0.1795</td>
<td>0.1115</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>book</td>
<td>buch</td>
<td>0.25</td>
<td>0.5</td>
<td>0.6667</td>
<td>0.7996</td>
<td>...</td>
<td>1</td>
</tr>
<tr>
<td>a</td>
<td>buch</td>
<td>0.25</td>
<td>0.25</td>
<td>0.1538</td>
<td>0.0889</td>
<td>...</td>
<td>0</td>
</tr>
<tr>
<td>book</td>
<td>ein</td>
<td>0.25</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>...</td>
<td>0.5</td>
</tr>
<tr>
<td>a</td>
<td>ein</td>
<td>0.25</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>...</td>
<td>0.5</td>
</tr>
<tr>
<td>the</td>
<td>haus</td>
<td>0.25</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>...</td>
<td>0.5</td>
</tr>
<tr>
<td>house</td>
<td>haus</td>
<td>0.25</td>
<td>0.5</td>
<td>0.5</td>
<td>0.5</td>
<td>...</td>
<td>0.5</td>
</tr>
</tbody>
</table>
Other word-based translation and alignment models

An incomplete sampling of other work in this area
Word-based translation models review

- Introduced probabilistic models of the form

\[ P(e|f) = \sum_a P(e, a|f) \]

- Used hidden alignment to explain the generation of target given source

- This lecture
  - Other models of the same type (extensions)
  - Discriminative word – alignment models trying to derive the correct word-level alignment to match a gold standard
    - \( P(a|e, f) \)
Extensions to HMM word-based translation model

- Toutanova et al 02 EMNLP
  - Limited notion of fertility (as allowed by the independence assumptions)
  - Word insertion depends on other target words
    - Intuition: usually function words are inserted and they should be justified by content target words

- Xiaodong He 07, 2nd SMT workshop
  - Word-dependent model of distortion with smoothing
  - Outperforms IBM-4 on AER
  - Outperforms IBM-4 when used as alignment for phrase-based translation on Europal data
  - More than 4 times faster than IBM-4
A Generative Model handling many-to-many alignments: LEAF

- The LEAF generative model (Fraser & Marcu 2007) allows many-to-many alignments between source and target.
- The groups of words in the source and target that are aligned can be non-consecutive.
- The model is further improved by semi-supervised learning: using a small amount of labeled aligned data.
- Improves alignment F-measure and BLEU relative to unsupervised and semi-supervised IBM Model 4.

<table>
<thead>
<tr>
<th>System</th>
<th>French/English F-Measure ($\alpha = 0.4$)</th>
<th>BLEU</th>
<th>Arabic/English F-Measure ($\alpha = 0.1$)</th>
<th>BLEU</th>
</tr>
</thead>
<tbody>
<tr>
<td>GIZA++</td>
<td>73.5</td>
<td>30.63</td>
<td>75.8</td>
<td>51.55</td>
</tr>
<tr>
<td>(FRASER AND MARCU, 2006b)</td>
<td>74.1</td>
<td>31.40</td>
<td>79.1</td>
<td>52.89</td>
</tr>
<tr>
<td>LEAF UNSUPERVISED</td>
<td>74.5</td>
<td></td>
<td>72.3</td>
<td></td>
</tr>
<tr>
<td>LEAF SEMI-SUPERVISED</td>
<td>76.3</td>
<td>31.86</td>
<td>84.5</td>
<td>54.34</td>
</tr>
</tbody>
</table>
Agreement

- Most word-based translation models we have seen so far are asymmetric
  - One language is source, another is target; implications about the directionality of allowed alignments

- [Liang et al 06] Alignment by agreement.
  - Learns HMM models in both directions
  - Adds a term to the log-likelihood that encourages agreement between the models in the two directions

- [Graca et al 08] Incorporating agreement constraints using posterior regularization.
Alignment by Agreement [Liang et al 2006]

Two complementary models

Fractional alignments

Slide by Percy Liang
Discriminative word alignment

- Make use of parallel sentences annotated with gold-standard alignments
  - turns into a structured prediction problem on which we can use standard machine learning approaches
  - Data publicly available for English-French, English-Chinese, English-Arabic, English-Romanian and possibly other languages

- General approaches
  - Use generative unsupervised word-based models as a source of features
  - Can use multiple overlapping features more easily, no need to come up with a generative story
  - Inference and training can still be very expensive if we want to model alignment dependencies well -> work on approximate inference, new algorithms
Discriminative word alignment

- Moore et al. 06 [using a perceptron]
  - Staged training and collection of statistics over large un-annotated data
  - Approximate inference algorithm
- Taskar et al. 05, Cherry and Lin 06, Lacoste-Julien et al. 06 [SVM]
- Blunsom & Cohn 06 [CRF]
- Haghighi et al. 08
  - Uses a block ITG grammar to define the space of possible alignments
  - Better in AER and BLEU compared to HMM and IBM-4
  - Code available
Using linguistic information to improve word alignment

- POS tags used to condition distortions
- Syntactic constituents used to define constraints on alignments in discriminative models (Cherry & Lin 06)
- Using morphological information
  - Goldwater and McClosky 05
    - Report translation results using word-based translation models with different morphological pre-processing
  - Popović and Ney 04
    - Incorporate morpho-syntactic information in IBM models
  - Fraser and Marcu 05
    - Evaluate effect of stemming for Romanian-English alignment
Using linguistic information to improve word alignment

- Simultaneous morpheme segmentation and morpheme alignment using linguistic features [Naradowsky and Toutanova 2011]
Inducing word-translation lexicons without parallel corpora

- Many references in book
- Rapp 95
  - Perhaps the first work in this area, uses similarity of co-occurrence vectors
- Koehn & Knight 00
  - Uses a lexicon (without probabilities) and uses monolingual text to estimate probabilities
- Koehn & Knight 02
  - Extensions to co-occurrence model
- Garera et al 09
  - Use dependency analyses, match based on POS
- Haghighi et al 08
  - Uses co-occurrence and orthographic features and CCA (canonical correlation analysis) to estimate a matching
Phrase-based translation models
Motivation

- Word-based translation models condition target words only on their aligned source word
- Too strong an assumption
  - Especially for one-to-many correspondences
  - For inserted and deleted words
  - In general, much better if the model can condition on more source/target context
- Restrictions on the alignment space are too strong in some cases
  - One-to-many in some direction (need many-to-many)
Basic phrase-translation overview

- Decisions for target sentence, segmentation and alignment, given source sentence

- Source sentence is segmented into source phrases
  - Not linguistically motivated segmentation
  - Segmentation distribution not modeled

- Each source phrase is translated into a target phrase
  - Independent of other source phrases and their translations

- The resulting target phrases are re-ordered to form output
Generative model notation

- Segmentation notation
  - Segmenting a sentence into $l$ phrases, each of which is denoted by $\overline{e_i}$
    \[ e, S = [\overline{e_1}, \overline{e_2}, ..., \overline{e_l}] = \overline{e_I} \]

- We will use the noisy channel formulation, so generating source $f$ given target $e$

- $f$ is also segmented into corresponding phrases and reordered
  \[ f, S_f = [\overline{f_{a_1}}, \overline{f_{a_2}}, \overline{f_{a_I}}] \]

- $\overline{f_i}$ is the foreign phrase aligned to source phrase $\overline{e_i}$

- $\overline{f_{a_i}}$ is the foreign phrase in foreign position $i$

- Differs a bit from textbook where there is no notation for the target order
Phrase translation model

Distortion model
\[ d(x) = \alpha^{|x|} \]

Not a normalized probability distribution
Combining with language model

- \( P(e, A, S_f, S_e | f) \sim \propto P_{LM}(e)P(f, A, S_f, S_e | e) \)
  \[ \sim P_{LM}(e) \prod_{i=1}^{I} \phi(f_i | e_i)\alpha^{\text{start}_i-\text{end}_{i-1}-1} \]

- \( P(e | f) \sim \propto P_{LM}(e) \max_{A, S_f, S_e} P(f, A, S_f, S_e | e) \)

- \( P_{LM}(\text{Tomorrow I will fly to the conference in Canada}) \)
- \( \phi(\text{Morgen}|\text{Tomorrow})\phi(\text{ich}|I)\phi(\text{fliege}|\text{will fly}) \ldots \)
- \( \alpha^0 \alpha^1 \alpha^2 \ldots \)
Differences from word-based translation models

- Basic unit for translation probabilities is a phrase, not word
- Alignment between source and target phrases is one-to-one (not one-to-many)
  - But the word-level correspondences within phrases can be many-to-many
- No insertion or deletion of phrases in the basic model
  - Deletion and insertion of words within the context of other words in phrases
- Translation probabilities are not estimated using ML estimation of incomplete data, but using heuristics and word-based models
  - Why: it is easy and it works
  - For principled estimation to work we need better models; starting to see some success (more later)
Learning phrase translation pairs and their probabilities

- Train word-based translation (alignment models)
- Align parallel sentence pairs in training data
- Extract all phrase-pairs consistent with word alignment
- Estimate phrase translation probabilities from counts in aligned training data
  - Can use word-based models for smoothing
Extracting phrase pairs

- Start with *word-aligned* sentences

- Extract phrase-pairs (up to some length) that are *consistent* with the word alignment
Which phrases are consistent with a word alignment?

- Depends how we are going to use the phrases
- In the current system, each source phrase is paired with exactly one target phrase
  - Once we translate a source phrase, we cannot re-use it again to add something to the translation
    - The target side of each phrase pair should contain the complete translation of the source side
  - Once we generate a target phrase from a given source phrase, we cannot add additional source phrases as an explanation of the target phrase
    - The target side of each phrase pair should contain no more material than is warranted by the source side
- Look for source-target phrase pairs which are translationally equivalent in some context (hopefully, many contexts)
Phrases consistent with alignment

A phrase pair is consistent with alignment $A$ iff:

$$\forall e_i \in \bar{e} : (e_i, f_j) \in A \rightarrow f_j \in \bar{f}$$

AND $$\forall f_j \in \bar{f} : (e_i, f_j) \in A \rightarrow e_i \in \bar{e}$$

AND $$\exists e_i \in \bar{e} , f_j \in \bar{f} : (e_i, f_j) \in A$$
Word alignment induced phrases (1)

(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green)
Word alignment induced phrases (2)

(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green),
(Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the),
(bruja verde, green witch)
Word alignment induced phrases (3)

(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green),
(Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the),
(bruja verde, green witch), (Maria no daba una bofetada, Mary did not slap),
(no daba una bofetada a la, did not slap the), (a la bruja verde, the green witch)
Word alignment induced phrases (4)

(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green),
(Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the),
(bruja verde, green witch), (Maria no daba una bofetada, Mary did not slap),
(no daba una bofetada a la, did not slap the), (a la bruja verde, the green witch),
(Maria no daba una bofetada a la, Mary did not slap the),
(daba una bofetada a la bruja verde, slap the green witch)
(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green),
(Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the),
(bruja verde, green witch),  (Maria no daba una bofetada, Mary did not slap),
(no daba una bofetada a la, did not slap the), (a la bruja verde, the green witch),
(Maria no daba una bofetada a la, Mary did not slap the), (daba una bofetada a la bruja verde, slap the green witch),  (no daba una bofetada a la bruja verde, did not slap the green witch),
(Maria no daba una bofetada a la bruja verde, Mary did not slap the green witch)
Phrase extraction and null-aligned words

michael — michael
michael assumes — michael geht davon aus ; michael geht davon aus ,
michael assumes that — michael geht davon aus , dass

assumes — geht davon aus ; geht davon aus ,
assumes that — geht davon aus , dass

that — dass ; , dass
Estimating phrase-translation probabilities

- Estimation using relative frequency
- Assuming every phrase pair occurs as many times as there are sentences from which we have extracted the phrase pair

\[
\phi(f_i | e_i) = \frac{\text{count}(f_i, e_i)}{\sum_{f_i'} \text{count}(f_i', e_i)}
\]

- Does not make sense as a generative model because it assumes source and target phrases are generated multiple times
- Works well and hard to beat with more principled approaches
Extensions to basic phrase-translation model
So far we almost have a probabilistic model
- Even though estimation is not principled and the distortion model is not normalized

The model still makes strong independence assumptions

We can do better by using a generative-discriminative hybrid model
- The current generative components: phrase translation, distortion, language model can be features (log-probabilities)
- We can learn weights for them discriminatively (to optimize translation performance, e.g. BLEU)

\[
\text{score}(e, f, A, S_e, S_f) = \lambda_1 \log(P_{LM}(e)) + \lambda_2 \log P_{TM}(f, A, S_f, S_e | e) + \lambda_3 \text{dist}(e, f, A, S_e, S_f)
\]
Translation using log-linear model

- The translation $e$ of $f$ is given by
  \[
  \arg \max_{e, A, S_e, S_f} \text{score}(e, f, A, S_e, S_f)
  \]

  \[
  \text{score}(e, f, A, S_e, S_f) = \lambda_1 \log(P_{LM}(e))
  \]

  \[
  + \lambda_2 \log(P_{TM}(f|e) \cdot P_{TM}(A, S_f|S_e))
  \]

  \[
  + \lambda_3 \text{dist}(e, f, A, S_e, S_f)
  \]

- Just by fitting separate weights for the three components we can do much better in BLEU
  - The basic model is equivalent to this model using all weights $= 1$

- We can also add other features, without worrying about the generative story
Additional features in log-linear translation model

- Phrase translation probabilities in the other direction as well
  - \( \phi(\bar{e}_i|\bar{f}_i) \) (estimated using counts as the other direction)
- Could devise a more complex lexicalized sub-model for the reordering decisions
- Number of phrase pairs / used
- Number of words in the target sentence
- Other phrase-translation models for smoothing (Lexical weighting)
- Other language models, new features you come up with for your course project?
Word count and phrase count features

- Word count: how many words does the output sentence have
  \[= |e| \]
- Not modeled explicitly so far
- Language model prefers shorter sentences
- The BLEU score does not like sentences that are too short
  - Depending on how well the model is doing, this feature can help it come up with a tradeoff between precision and brevity so as to maximize BLEU
- Phrase count: the count \(l\) of phrase-pairs used
  - Smaller number of phrases is preferred by phrase-translation model
  - But sometimes a larger number of smaller phrases is better: estimated more robustly
Lexical weighting feature

- Assigns conditional probability to target phrase given source phrase using translation probabilities from word-based models and fixed alignment.

\[
\frac{1}{3} \left[ w(\text{assumes}|\text{geht}) + w(\text{assumes}|\text{davon}) + w(\text{assumes}|\text{aus}) \right] 
\]

- Helps derive more robust estimates in case of sparse data.
- Also used in both directions.
Lexicalized re-ordering model

- The only explicit model of re-ordering so far looks only at distortion in the source sentence
- Can have a model looking at more information: e.g. words from source and target phrases and other information

- Simple lexicalized model
  - for each phrase pair $[\bar{e}_i, \bar{f}_i]$, classify its re-ordering pattern with respect to the previous pair $[\bar{e}_{i-1}, \bar{f}_{i-1}]$ in three types
    - (m) monotone order (d=0)
    - (s) swap with previous phrase
    - (d) discontinuous
Orientation of phrase pairs

Phrase pair 1: monotone
Phrase pair 2: discontinuous
Phrase pair 3: swap
Phrase pair 4: discontinuous
Predict orientation given phrase-pair

- \( P_o(orientation | \bar{s}, \bar{f}) \)

- Collect counts of orientation types given phrase pair in word-aligned parallel data, relative frequency with smoothing
  - Alignment point at top left = monotone
  - Alignment point at top right = swap
  - Otherwise discontinuous

[from Koehn et al., 2005, IWSLT]
Lexicalized reordering feature

- To compute the feature value for a full translation hypothesis
  - Log of product of orientation probabilities of individual phrase pairs
    
    $$ h_{lo}(e, f, S_e, S_f, A) = \sum_{i} \log P_o(\text{orient}_i | \overline{f}_i, \overline{e}_i) $$

- A deficient probabilistic model of distortion
  - Assumes orientations are independent which leads to impossible configurations

- Still very powerful when used as a feature function

- Other phrase-based re-ordering models proposed in literature, with similar gains in performance
Impact of lexicalized reordering

- Figure from Koehn IWSLT 2005, also investigating variations in exact definition of reordering events

<table>
<thead>
<tr>
<th>Language Pair</th>
<th>Best Lexicalised Reordering</th>
<th>Word Alignment</th>
<th>Baseline</th>
<th>Improved</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic-English</td>
<td>orientation-bidirectional-fe</td>
<td>final-and</td>
<td>49.9</td>
<td>50.9</td>
</tr>
<tr>
<td>Japanese-English</td>
<td>orientation-fe</td>
<td>intersect</td>
<td>45.1</td>
<td>47.6</td>
</tr>
<tr>
<td>Korean-English</td>
<td>orientation-fe</td>
<td>intersect</td>
<td>35.7</td>
<td>42.3</td>
</tr>
<tr>
<td>Chinese-English</td>
<td>monotonicity-fe</td>
<td>intersect</td>
<td>34.6</td>
<td>38.6</td>
</tr>
<tr>
<td>English-Chinese</td>
<td>monotonicity-bidirectional-fe</td>
<td>grow-diag</td>
<td>15.2</td>
<td>16.6</td>
</tr>
</tbody>
</table>

Table 3: Best lexicalised reordering methods, compared against the baseline (using only distance-based reordering penalty): Improvements for all language pairs
Fitting log-linear translation model weights

- Approximate search to maximize BLEU score of resulting model
- Split data into training, development, and test sets
- Train word-alignment model and extract phrases from training set
- Fit the weights of feature functions in log-linear model on dev set, to maximize BLEU
  - Iteratively generate N-best lists of translation hypotheses
  - Adjust parameters to move better translations to top
Discriminative training

More in Chapter 9. [Och 2003]
Discriminative versus generative model

Results from Och & Ney 02 using a slightly different framework [alignment templates].

Weights have not been trained to maximize BLEU but to maximize log-likelihood of log-linear model.
Effect of discriminative training

<table>
<thead>
<tr>
<th>error criterion used in training</th>
<th>mWER [%]</th>
<th>mPER [%]</th>
<th>BLEU [%]</th>
<th>NIST</th>
<th># words</th>
</tr>
</thead>
<tbody>
<tr>
<td>confidence intervals</td>
<td>+/- 2.7</td>
<td>+/- 1.9</td>
<td>+/- 0.8</td>
<td>+/- 0.12</td>
<td>-</td>
</tr>
<tr>
<td>MMI</td>
<td>68.0</td>
<td>51.0</td>
<td>11.3</td>
<td>5.76</td>
<td>21933</td>
</tr>
<tr>
<td>mWER</td>
<td>68.3</td>
<td>50.2</td>
<td>13.5</td>
<td>6.28</td>
<td>22914</td>
</tr>
<tr>
<td>smoothed-mWER</td>
<td>68.2</td>
<td>50.2</td>
<td>13.2</td>
<td>6.27</td>
<td>22902</td>
</tr>
<tr>
<td>mPER</td>
<td>70.2</td>
<td>49.8</td>
<td>15.2</td>
<td>6.71</td>
<td>24399</td>
</tr>
<tr>
<td>smoothed-mPER</td>
<td>70.0</td>
<td>49.7</td>
<td>15.2</td>
<td>6.69</td>
<td>24198</td>
</tr>
<tr>
<td>BLEU</td>
<td>76.1</td>
<td>53.2</td>
<td>17.2</td>
<td>6.66</td>
<td>28002</td>
</tr>
<tr>
<td>NIST</td>
<td>73.3</td>
<td>51.5</td>
<td>16.4</td>
<td><strong>6.80</strong></td>
<td>26602</td>
</tr>
</tbody>
</table>

Table from Och 03.
More principled probabilistic models for phrase-based SMT
What we have seen so far

- Phrase translation generative model where we estimate the phrase-translation probabilities heuristically
  - Using counts of phrase pairs extracted from word-aligned data

- Try a more principled approach
  - Define a probabilistic generative model of target and source sentences (or target given source)
  - The model will use hidden variables for segmentation and alignment between phrases
  - We can estimate the model parameters from incomplete data
    - Maximum Likelihood
    - Maximum a posteriori (if we have a prior)
    - Fully Bayesian inference (marginalize over model parameters)
A Joint Model for Phrasal Alignment

[Marcu and Wong 2002]

- Generate source and target sentences $f, e$ jointly using a decomposition into concepts
  - Chose number of phrase pairs (concepts) to generate
- Generate the phrase pairs $\overline{f}_i \overline{e}_i$ in source order
- Place each target phrase $\overline{e}_i$ in position $pos_i$

$$P(e, f, A, S_e, S_f) = \prod_{i} t(\overline{f}_i, \overline{e}_i) d(pos_i | pos_{i-1})$$
Complexity of Alignment and Segmentation Space for Joint Model

- Number of ways to segment sentence $f$ into $m$ contiguous phrases if $\text{len}(f)=n$
  - $n$ choose $m$
- Similarly for segmenting the source
- Multiplied by number of 1-to-1 alignments b/w phrases
  - $m!$
- How many possible phrase-pairs should be considered
  - $O(n^4)$
- = too large to sum over exhaustively
- Various approximations for speeding it up
- Pruning of possible phrase-pairs using frequency cutoffs
- Results in translation better than IBM-4
Other extensions

- Constrain the alignment space for the joint model using best alignments from word-based models [Birch et al]
- Constrain the re-ordering space using an ITG model [Cherry an Lin 2007]
  - results about equal to the heuristic estimation approach
- Problem for conditional models and likelihood training
  - Prefer to make source phrases as large as possible (can get probability 1 for the training data) [DeNero et al 06]
- Use a Bayesian model which prefers short phrases and also prefers consistency with word-based alignment models [DeNero et al 08]
  - Devise operators for sampling using this model (operators in Gibbs sampling)
## Results from DeNero et al 08

<table>
<thead>
<tr>
<th>Estimate</th>
<th>Phrase Pair Count</th>
<th>NIST BLEU</th>
<th>Exact Match METEOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heuristic</td>
<td>4.4M</td>
<td>29.8</td>
<td>52.4</td>
</tr>
<tr>
<td>DP</td>
<td>0.6M</td>
<td>28.8</td>
<td>51.7</td>
</tr>
<tr>
<td>HDP</td>
<td>0.3M</td>
<td>29.1</td>
<td>52.0</td>
</tr>
<tr>
<td>DP-composed</td>
<td>3.7M</td>
<td>30.1</td>
<td>52.7</td>
</tr>
<tr>
<td>HDP-composed</td>
<td>3.1M</td>
<td>30.1</td>
<td>52.6</td>
</tr>
<tr>
<td>DP-smooth</td>
<td>4.8M</td>
<td>30.1</td>
<td>52.5</td>
</tr>
<tr>
<td>HDP-smooth</td>
<td>4.6M</td>
<td><strong>30.2</strong></td>
<td>52.7</td>
</tr>
<tr>
<td>Heuristic + lex</td>
<td>4.4M</td>
<td>30.5</td>
<td>52.9</td>
</tr>
<tr>
<td>DP-smooth + lex</td>
<td>4.8M</td>
<td>30.4</td>
<td>53.0</td>
</tr>
<tr>
<td>HDP-smooth + lex</td>
<td>4.6M</td>
<td><strong>30.7</strong></td>
<td>53.2</td>
</tr>
</tbody>
</table>

Can achieve slightly better results compared to heuristic model.
Summary

- Other word-based translation and alignment models
  - Extensions to HMM word-based models
  - Agreement
  - Discriminative word-alignment and symmetrization
  - Using linguistic information
  - Learning translation lexicons from non-parallel data

- Phrase-based translation model
  - Phrase extraction
  - Model Extensions
    - Other features and discriminative estimation
    - Order Model
  - Probabilistic Models for phrase-based translation
Assignments

- Reading for this week
  - Chapter 5

- Reading for next week
  - Chapter 6: Decoding

- Office hours next week?